



Quadram
Institute

Science ▾ Health ▾
Food ▾ Innovation

A primer on metabarcoding

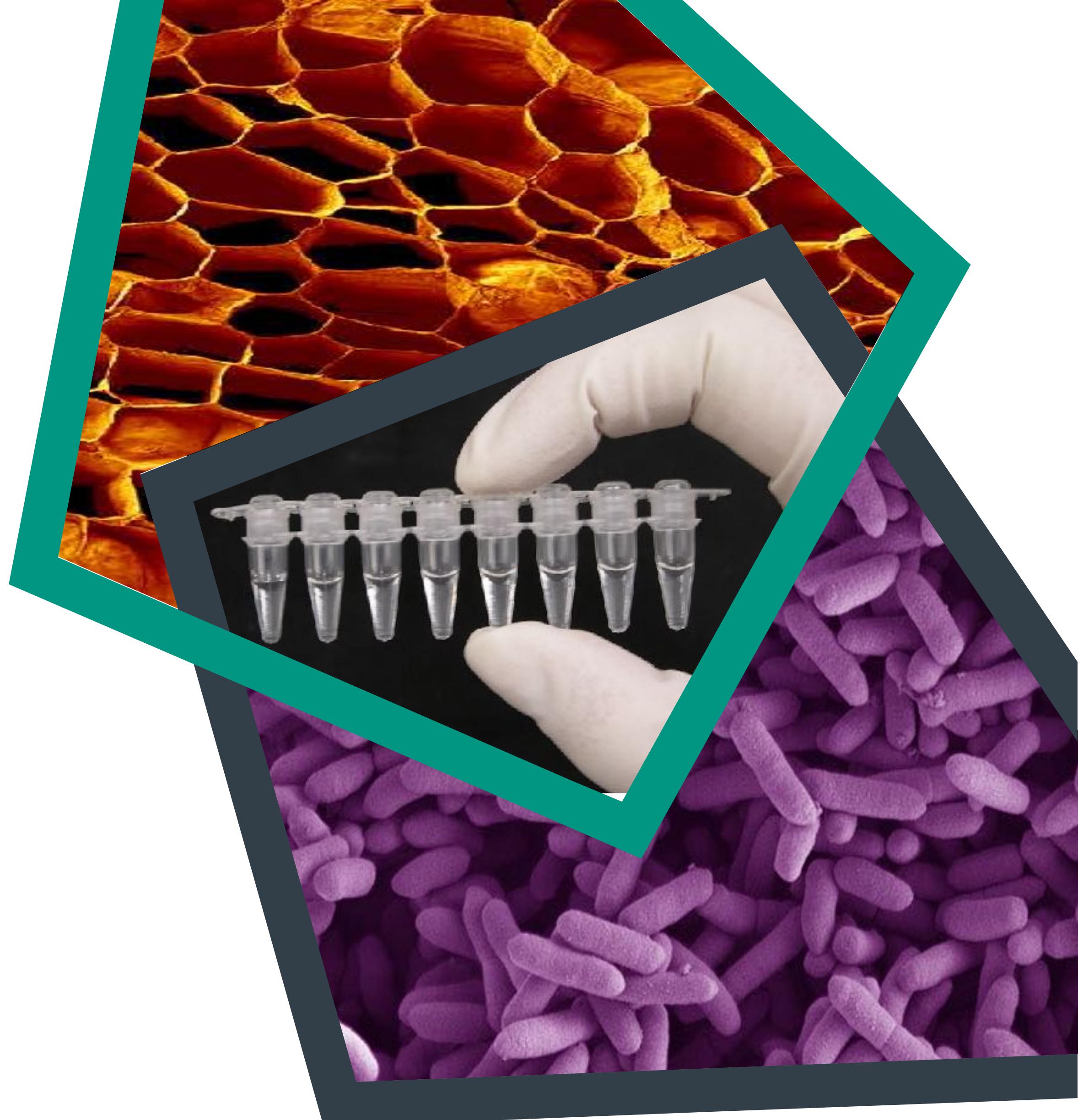
Andrea Telatin, Gut Microbial Health
Quadram Institute, UK





Science ▾ Health ▾
Food ▾ Innovation

What is metabarcoding about?



Metabarcoding

“Gene sequences,
most commonly those encoding rRNAs,
provide a basis for
estimating microbial phylogenetic diversity
and generating taxonomic inventories of
[...] microbial populations.”

– Sogin *et. al* 2006

Questions

- Who is in the sample?
- What are they doing?



Questions

- Who is in the sample?

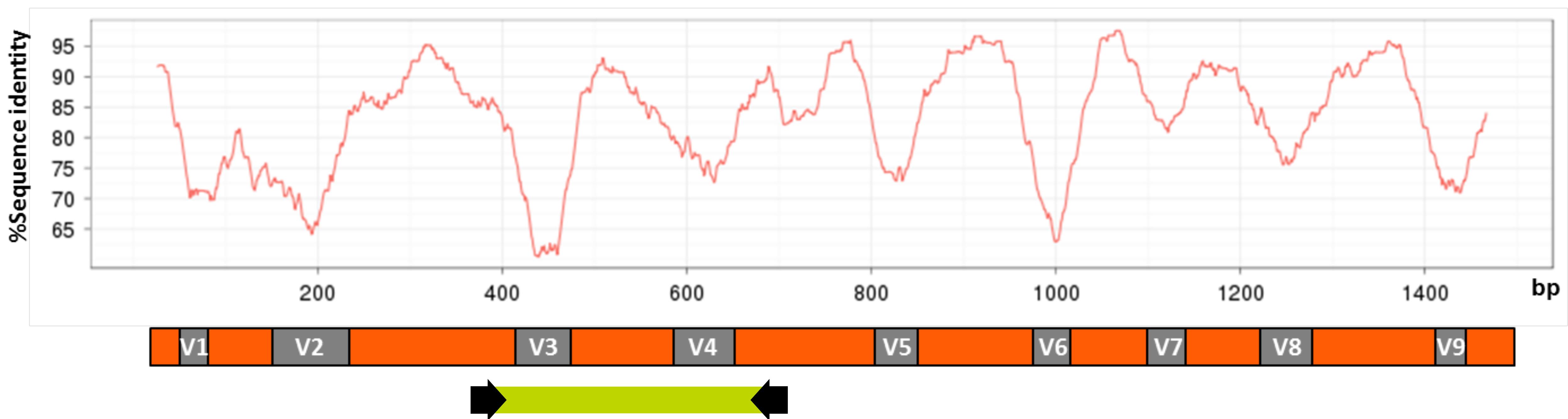
16S/ITS Amplicons, Metagenomics

- What are they doing?

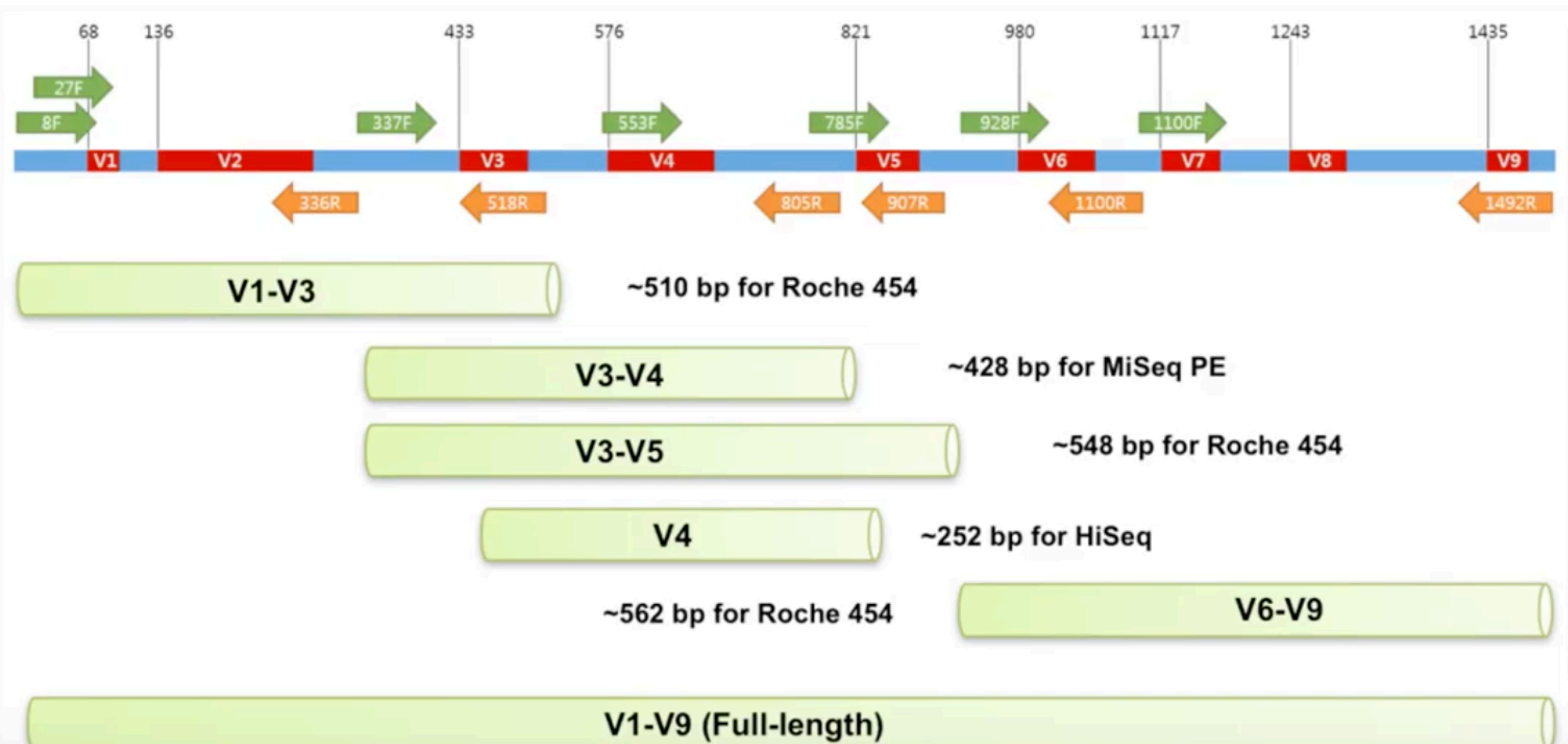
Metagenomics



Anatomy of a gene: 16S rDNA



Anatomy of a gene: 16S rDNA



Biases?

- Sampling (replicate or lie)
- Cell breaking (spores, cell wall,...)
- Amplify (primers? cycles? **chimeric molecules**)
- Sequencing
- Bioinformatics (assumptions, taxonomy databases...)

Biases?

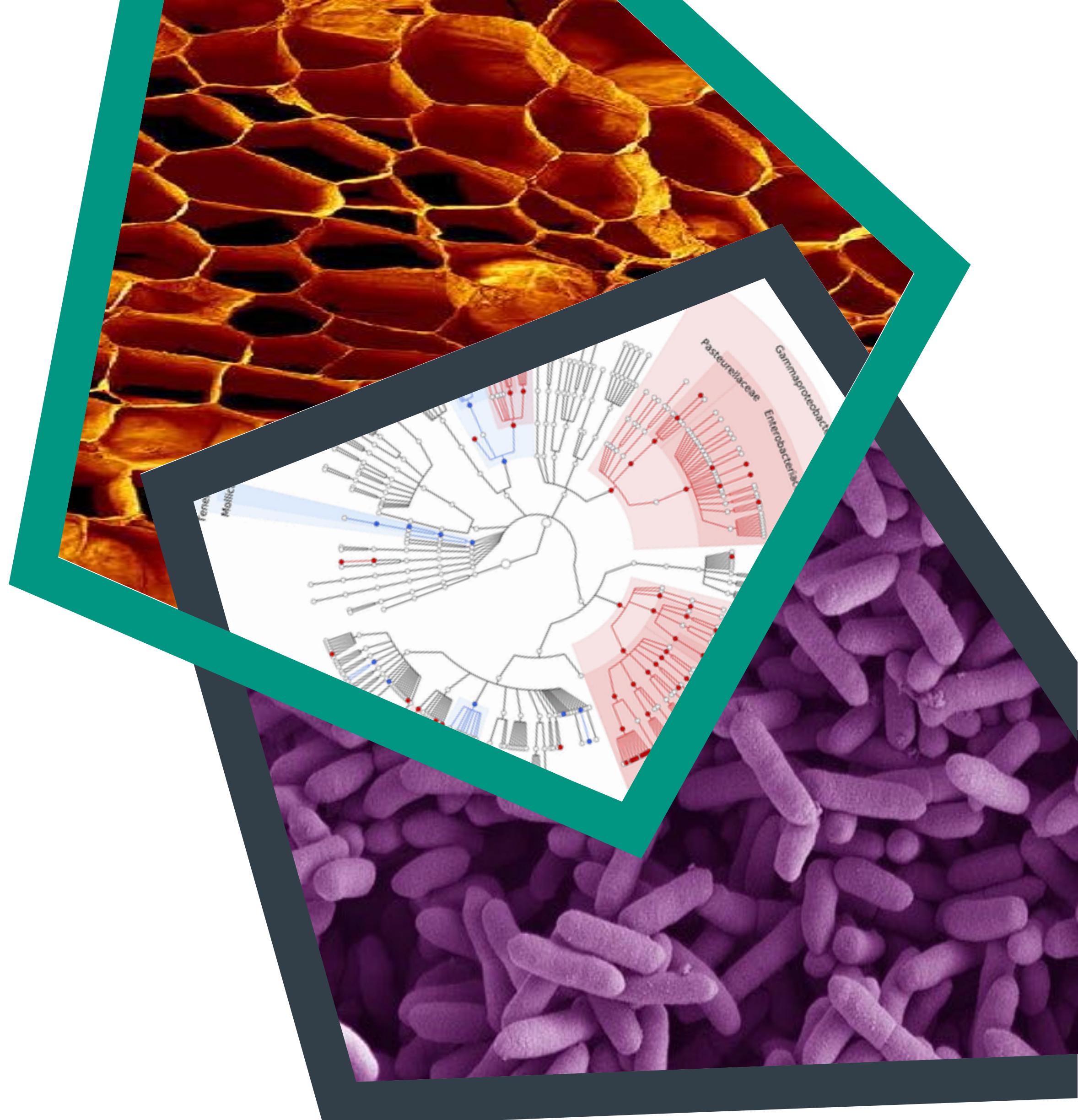
- Sampling (replicate or lie)
- Cell breaking (spores, cell wall,...)
- Amplify (primers? cycles? **chimeric molecules**)
- Sequencing
- Bioinformatics (assumptions, taxonomic assignment,...)

That's why we
compare different
samples...)

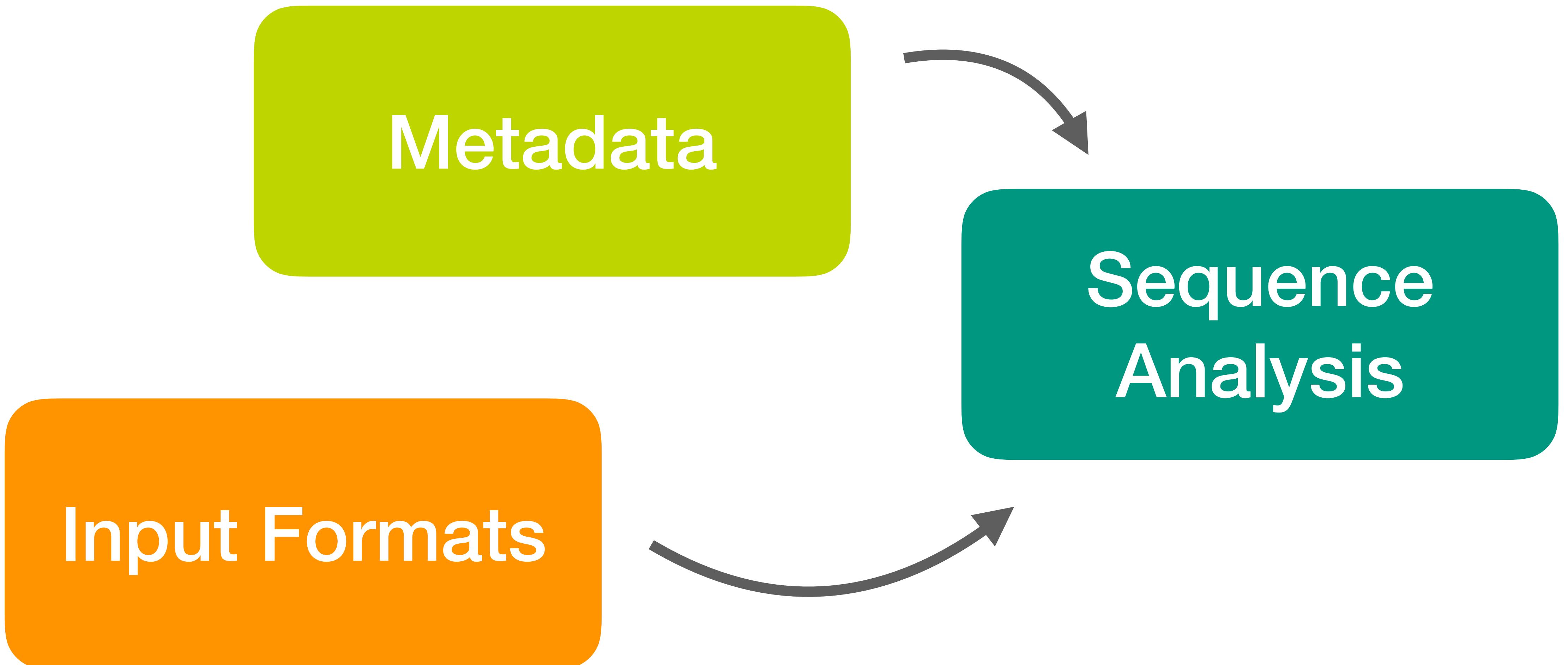


Science ▾ Health ▾
Food ▾ Innovation

Bioinformatics analysis



Tasks



How (tools)

QIIME 1.9

LOTUS

USEARCH

QIIME 2

MOTHUR

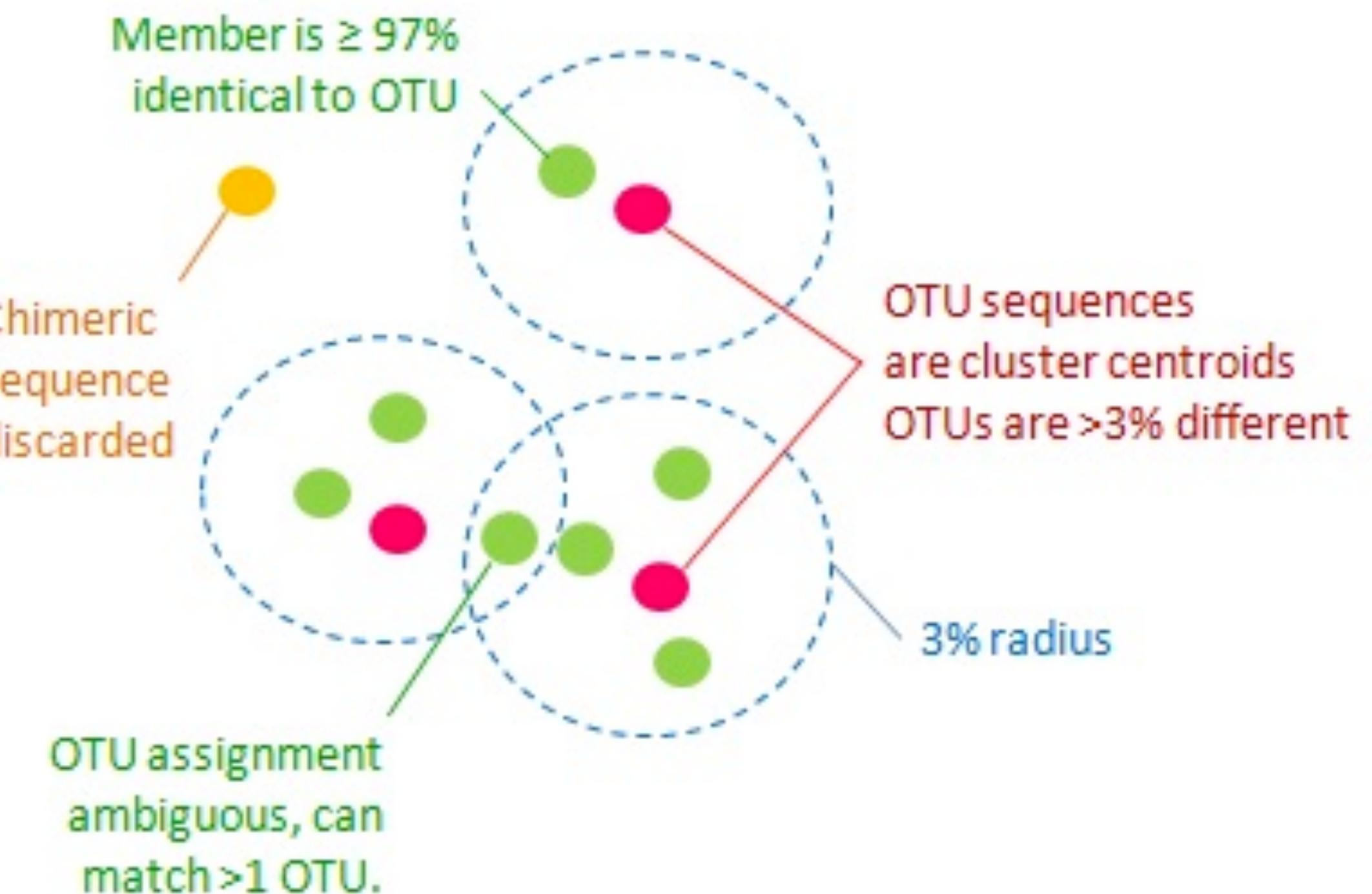
...

A training source: USEARCH

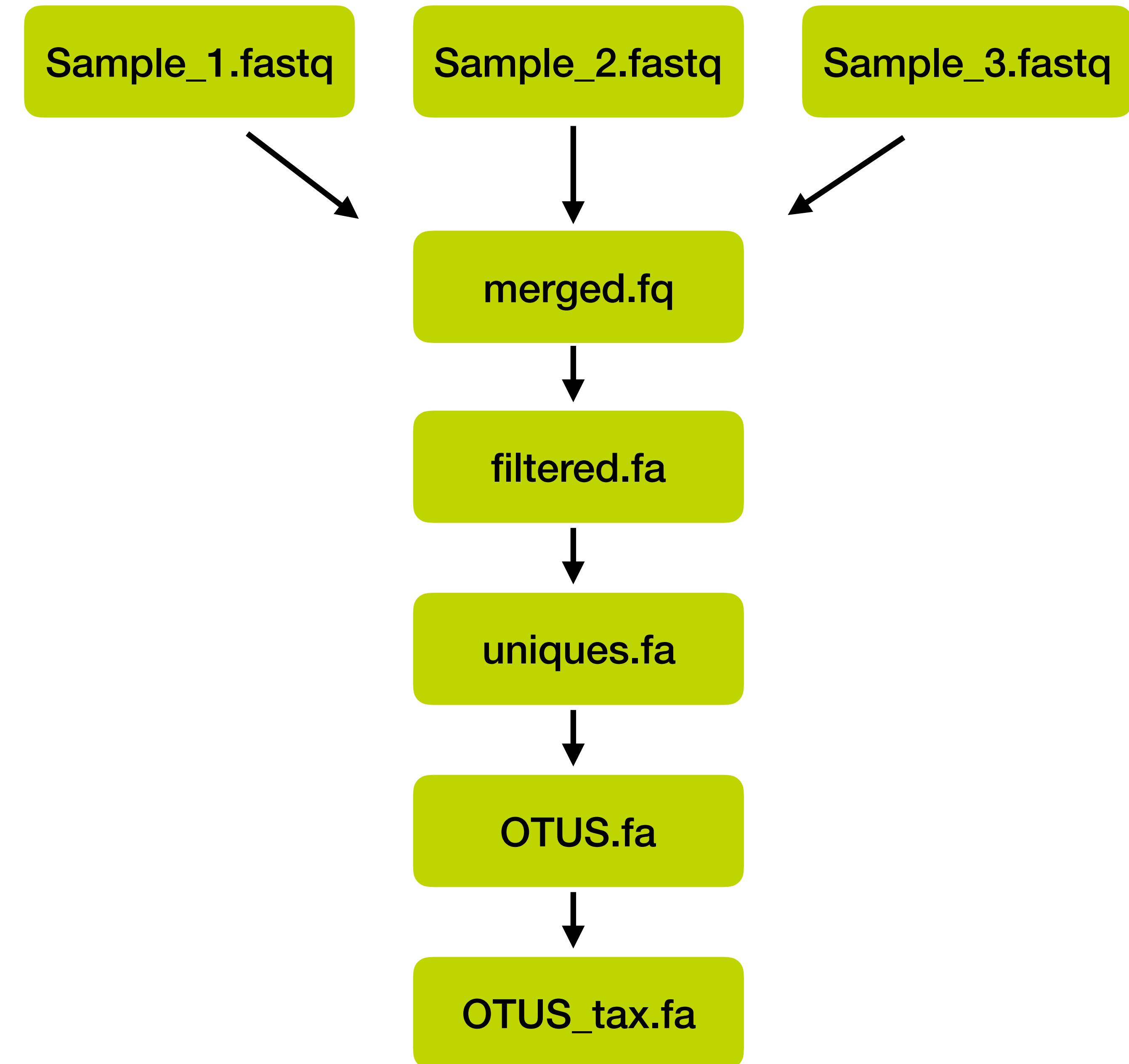
USEARCH is a monolithic package written by Robert Edgar, an independent bioinformatics researcher.

Can be used to see how we manipulate reads to make an annotated OTU table.

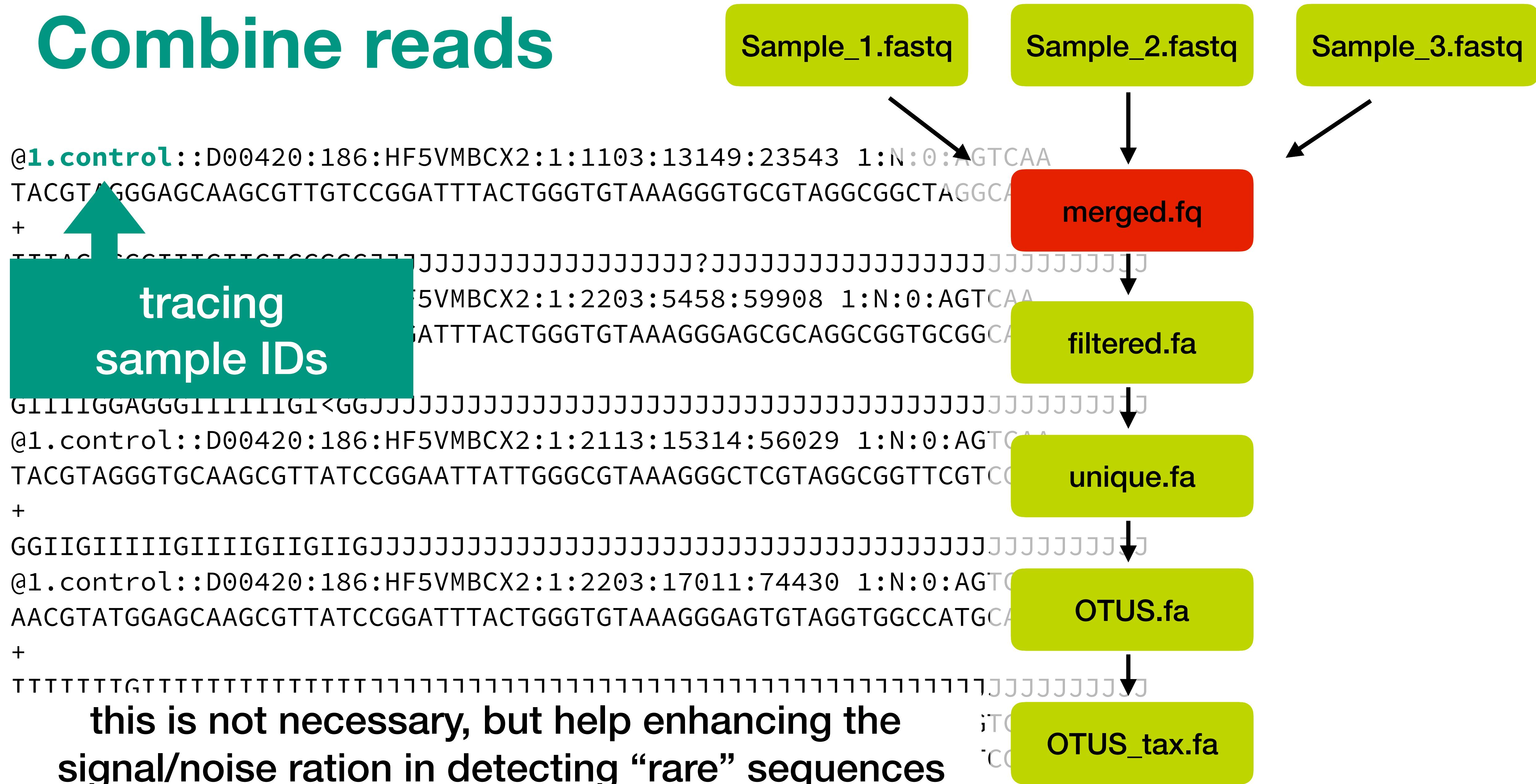
32-bit free download: <http://drive5.com/>



A simplified example



Combine reads



Quality filter

```
>Filt1
TACGTAGGGAGCAAGCGTTGCCGGATTTACTGGGTGTAAAGGGTGCCTAGGCGGCTAGGCAACTCACACGCTGAAACCA
TGGGCTAACCTGTGGACTGCGGTTGAAACTGTTAGCTTGAGTGAAGTAGAGGTAGGCAGGAAATGCGTAGAGATCGGGAGGAACACCAGTGGCGAAGGCAGCCTACTGGGCTTTAAGTACCGCTGAAACCATG
GGTAGCAAACAGGATTAGAAACCCCGTAGTCC
```

```
>Filt2
TACGTATGGTGCAAGCGTTATCCGGATTTACTGGGTGTAAAGGGAGCGCAGGCGGTGCCTAGGCAACTCACACGCTGAAACCA
GGGGCTAACCCCGGTACTGCATTGGAAACTGTCGTACTAGAGTGTGGAGGGGTAAGCGAAATTCCTAGTGTAGGGTG
AAATGCGTAGATATTAGGAGGAACACCAGTGGCGAAGGCAGCCTACTGGACGATAACTGACGCTGAGGOTCGAAAGCGTG
GGGAGCAAACAGGATTAGAAACCCCTGGTAGTCC
```

```
>Filt3
TACGTAGGGTGCAAGCGTTATCCGGAATTATTGGCGTAAAGGGCTCGTAGGCGGTCGTCGCTCGGTCCGTTGAAAGTCC
ATCGCTTAACGGTGGATCTGCGCCGGTACGGCGGGCTGGAGTGCCTAGGGGAGACTGGGAATTCCGGTGTAAACGGTG
GAATGTGTAGATATCGGGAAGAACACCAATGGCGAAGGCAGGTCTCTGGGCCGTTACTGACCGTG
GGGAGCGAACAGGATTAGAAACCCCCGTAGTCC
```

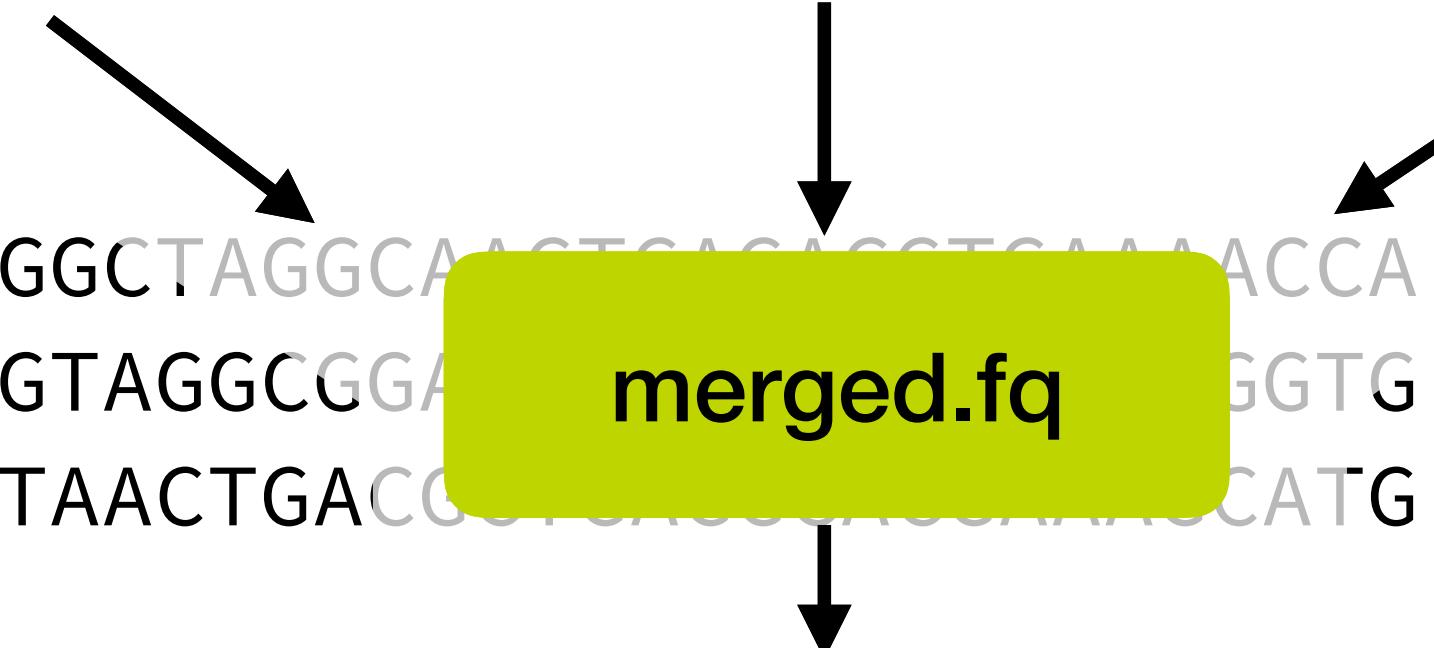
```
>Filt4
AACGTATGGAGCAAGCGTTATCCGGATTTACTGGGTGTAAAGGGAGTGTAGGTGGCCATGCAACTCACACGCTGAAACCATG
```

note that quality values are discarded

Sample_1.fastq

Sample_2.fastq

Sample_3.fastq



merged.fq

filtered.fa

unique.fa

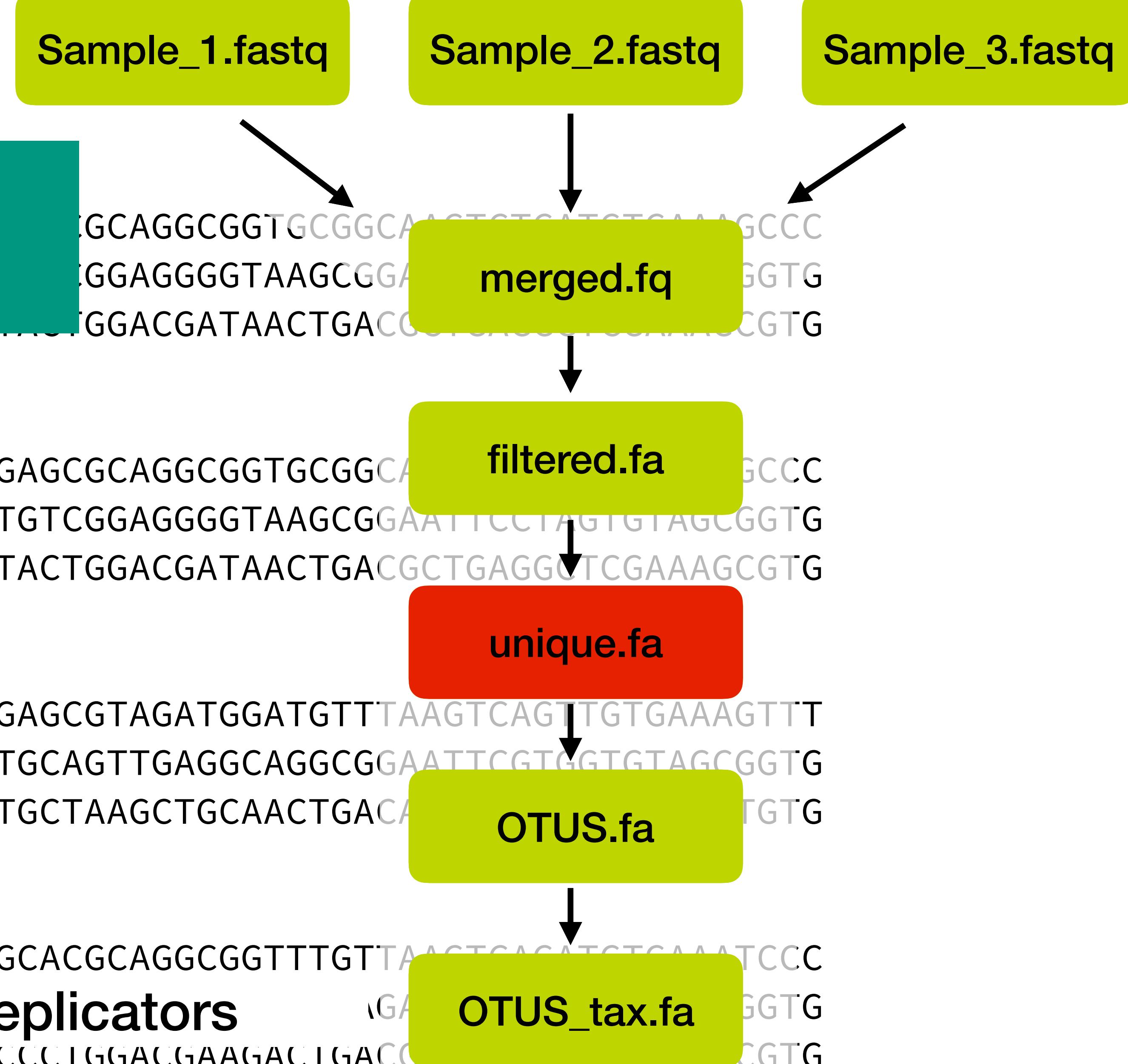
OTUS.fa

OTUS_tax.fa

Dereplication

```
>Uniq1;size=6103;  
TACGTATGGTGCAAGCGTTA  
GGGGCTAACCCCGGTACTGCA  
AAATGCGTAGATATTAGGAGGA  
GGGAGCAAACAGGATTAGAAACCCCCGTAGTCC  
>Uniq2;size=5788;  
TACGTATGGTGCAAGCGTTATCCGGATTTACTGGGTGTAAAGGGAGCGCAGGCGGTGC  
GGGGCTAACCCCGGTACTGCATTGGAAACTGTCGTACTAGAGTGTGGAGGGTAAGCG  
AAATGCGTAGATATTAGGAGGAACACCAGTGGCGAAGGCAGCTTACTGGACGATAACTG  
GGGAGCAAACAGGATTAGATAACCCCCGTAGTCC  
>Uniq3;size=5082;  
TACGGAGGATCCGAGCGTTATCCGGATTTATTGGGTTAAAGGGAGCGTAGATGGATGTT  
GCAGCTAACCGTAAAATTGCAGTTGATACTGGATATCTTGAGTGCAGTTGAGGCAGGC  
AAATGCTTAGATATCACGAAGAACTCCGATTGCGAAGGCAGCCTGCTAAGCTGCAACTG  
GGTATCAAACAGGATTAGAAACCCCCGTAGTCC  
>Uniq4;size=3438;  
TACGGAGGGTGCAAGCGTTAACCGGAATTACTGGCGTAAAGCGCACGCAGGCGTTGTT  
AAAIGCGTAGAGAGATCTGGAGGAAIACCGGIGGCAGGGCGCCCCCTGGACGAAGACTGAC
```

propagating
“metadata”



“sdm” and “usearch” are fast dereplicators

OTU/ASV picking

```
>0tu1
TACGTATGGTGCAAGCGTTATCCGGATTACTGGGTGTAAAGGGAGCGCAGGCGGTGC
GGGGCTCAACCCCCGGTACTGCATTGGAAACTGTCGTACTAGAGTGTGGAGGGGTAAGC
AAATGCGTAGATATTAGGAGGAACACCACTGGCGAAGGCGGCTTACTGGACGATAACTG
GGGAGCAAACAGGATTAGAAACCCCCGTAGTCC
```

```
>0tu2
TACGGAGGATCCGAGCGTTATCCGGATTATTGGGTTAAAGGGAGCGTAGATGGATGTT
GCGGCTCAACCGTAAAATTGCAGTTGATACTGGATATCTTGAGTGCAGTTGAGGCAGGC
AAATGCTTAGATATCACGAAGAACTCCGATTGCGAAGGCAGCCTGCTAACGCTGCAACTG
GGTATCAAACAGGATTAGAAACCCCCGTAGTCC
```

```
>0tu3
TACGGAGGGTGCAAGCGTTAACCGAATTACTGGCGTAAAGCGCACGCAGGCCGTTGTTAAGTC
CGGGCTCAACCTGGAACTGCATCTGATACTGGCAAGCTTGAGTCTCGTAGAGGGGGTAC
```

We can cluster sequences using identity (OTU),
or use sequencer-specific denoisers to detect high
quality sequence representatives (ASV)

```
ATCGCTTAACGGTGGATCTGCGCCGGGTACGGGCGGGCTGGAGTGCAGGTAGGGGAGACT
GAATGTGTAGATATCGGAAAGAACCCAATGGCGAAGGCAGGTCTGGGCCGTTACTGAC
```

Sample_1.fastq

Sample_2.fastq

Sample_3.fastq

merged.fq

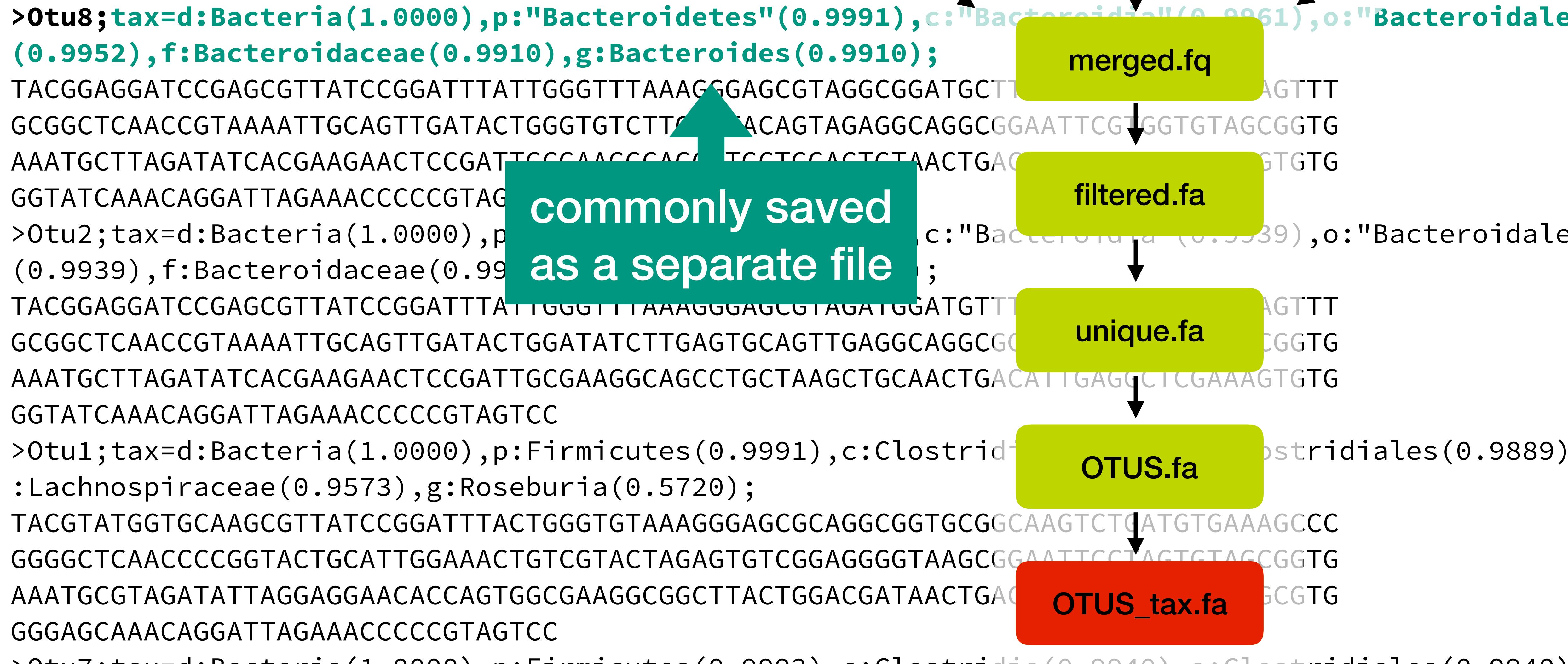
filtered.fa

unique.fa

OTUS.fa

OTUS_tax.fa

Classification



The OTU table

#OTU	S1	S2	S3	S4	S5
Otu1	14303	347	17513	474	14954
Otu2	10636	1312	7312	1486	11395
Otu4	3229	2662	7050	3612	3292
Otu3	382	9054	467	10925	400
Otu5	6523	1001	8289	1446	6546
Otu9	679	5963	1058	8398	682
Otu23	5025	413	6982	830	4976
Otu6	1213	6380	570	10908	1191
Otu7	2172	922	3836	1202	2278
Otu11	46	2386	70	2458	116

Produced counting how many reads of each sample map unambiguously against a specific OTU/ASV

Sample_1.fastq

Sample_2.fastq

Sample_3.fastq

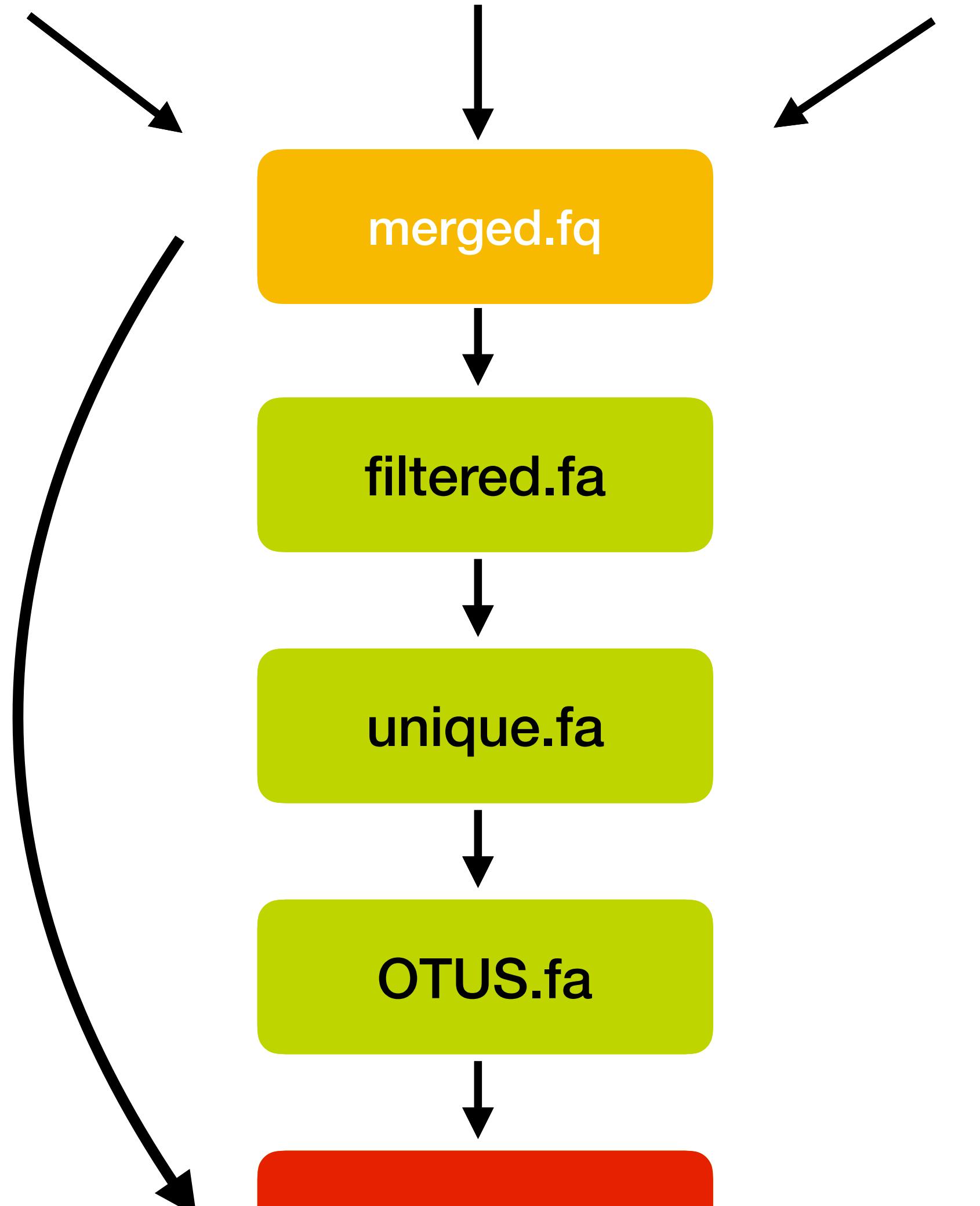
merged.fq

filtered.fa

unique.fa

OTUS.fa

OTUS_tax.fa



From the OTU table to numerical ecology

#OTU	S1	S2	S3	S4	S5	taxonomy
Otu1	14303	347	17513	474	14954	d:Bacteria;p:Firmicutes
Otu2	10636	1312	7312	1486	11395	d:Bacteria;p:Bacteroidetes;o:Bactero...
Otu4	3229	2662	7050	3612	3292	d:Bacteria;p:Proteobacteria;o:Alpha...
Otu3	382	9054	467	10925	400	d:Bacteria;p:Bacteroidetes
Otu5	6523	1001	8289	1446	6546	d:Bacteria;p:Firmicutes;o:Clostridia
Otu9	679	5963	1058	8398	682	d:Bacteria;p:Bacteroidetes;...
Otu23	5025	413	6982	83		
Otu6	1213	6380	570	1090		
Otu7	2172	922	3836	120		
Otu11	46	2386	70	245		
Otu8	3911	2136	3706	1997	4082	d:Bacteria;p:Proteobacteria;o:Gamma...

OTU table +
Taxonomy

	Diet	Sex	Date
S1	Normal	M	21-04
S2	Normal	F	22-04
S3	LowFat	F	21-04
S4	Low		5
S5	Big		4

Metadata
Table

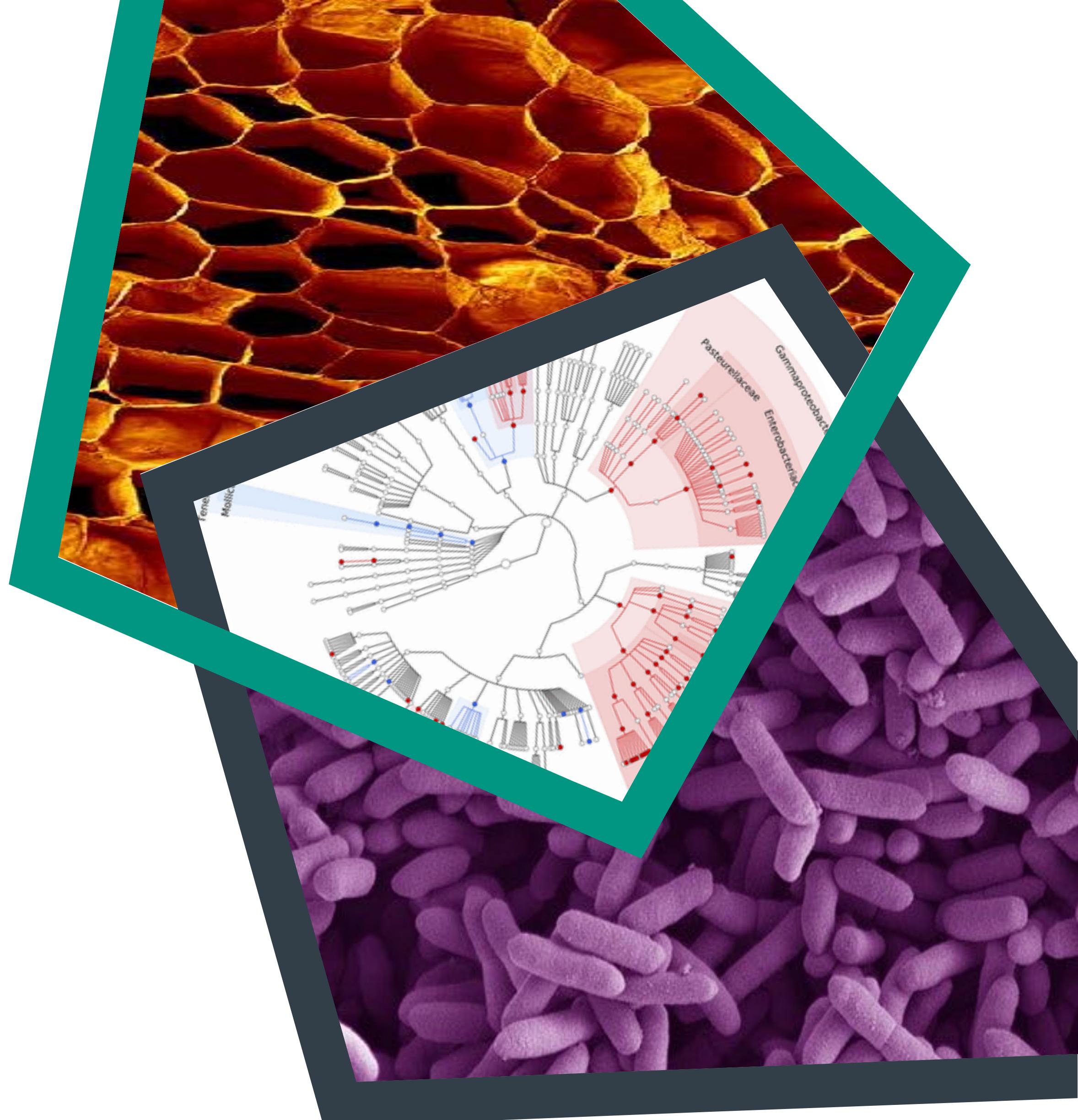


Normalization
Alpha Diversity
Beta Diversity
Differential OTU ab.
...



Science ▾ Health ▾
Food ▾ Innovation

Packages for metabarcoding analysis



Qiime 1.9

... an open-source bioinformatics pipeline to perform microbiome analysis from raw DNA sequencing data, now discontinued



- Rob Knight group and collaborators
- A collection of wrappers around well known tools (like UPARSE)
- The wrapping is **consistent** and makes **easier** to perform the whole analysis
- From sequences to **charts**
- It has been a very popular package but developed stopped years ago to produce a complete rewrite (Qiime 2)

Qiime 1.9

- With default parameters and database, was commonly overestimating diversity (i.e. several spurious OTUs identified)
- Now discontinued and should be avoided
- Still used for “compatibility” with legacy analyses

Metadata file (Mapping File in Qiime 1 jargon)

A TSV file linking Samples with barcodes (for demultiplexing) and other user defined columns

#SampleID	BarcodeSequence	LinkerPrimerSequence	ReversePrimer	Description
1	NNNNNNNNNNNN	CCCCGTCAATTCTTTRAGTTT	GTGCCAGCMGCCGCGGTAA	diseased
2	NNNNNNNNNNNN	CCCCGTCAATTCTTTRAGTTT	GTGCCAGCMGCCGCGGTAA	diseased
3	NNNNNNNNNNNN	CCCCGTCAATTCTTTRAGTTT	GTGCCAGCMGCCGCGGTAA	diseased
4	NNNNNNNNNNNN	CCCCGTCAATTCTTTRAGTTT	GTGCCAGCMGCCGCGGTAA	diseased
5	NNNNNNNNNNNN	CCCCGTCAATTCTTTRAGTTT	GTGCCAGCMGCCGCGGTAA	healthy
6	NNNNNNNNNNNN	CCCCGTCAATTCTTTRAGTTT	GTGCCAGCMGCCGCGGTAA	diseased
7	NNNNNNNNNNNN	CCCCGTCAATTCTTTRAGTTT	GTGCCAGCMGCCGCGGTAA	healthy
8	NNNNNNNNNNNN	CCCCGTCAATTCTTTRAGTTT	GTGCCAGCMGCCGCGGTAA	diseased
9	NNNNNNNNNNNN	CCCCGTCAATTCTTTRAGTTT	GTGCCAGCMGCCGCGGTAA	healthy
10	NNNNNNNNNNNN	CCCCGTCAATTCTTTRAGTTT	GTGCCAGCMGCCGCGGTAA	diseased
11	NNNNNNNNNNNN	CCCCGTCAATTCTTTRAGTTT	GTGCCAGCMGCCGCGGTAA	healthy
12	NNNNNNNNNNNN	CCCCGTCAATTCTTTRAGTTT	GTGCCAGCMGCCGCGGTAA	diseased
13	NNNNNNNNNNNN	CCCCGTCAATTCTTTRAGTTT	GTGCCAGCMGCCGCGGTAA	healthy
14	NNNNNNNNNNNN	CCCCGTCAATTCTTTRAGTTT	GTGCCAGCMGCCGCGGTAA	diseased

Qiime 2



- a complete rewrite of Qiime 1; introduce the new concept of **artifacts** (input/output packages with metadata)
- strongly advocates **DADA2** algorighm for **ASV** detection rather than OTU picking
- new release are produced frequently, unfortunately sometimes changing some commands (not a sign of a mature and stable package)
- Still producing charts, in my opinion sometimes worse than in the past
- Robust **traceability** and reproducibility at the expense of ease of use and interchange of data with other tools

Qiime2 artifacts

qza

Qiime2 archive

It's the output format of all Qiime2 programs. It's a ZIP files with both data and metadata.



qzv

Qiime2 visualization

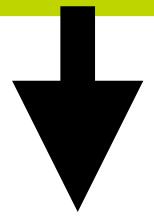
It's the output format for plots/charts and tables that the user could desire to inspect. It's an HTML document (web page) embedded in a “ZIP with metadata”, like qza.

<http://view.qiime2.org>

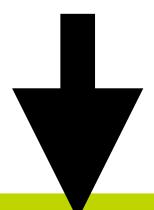
Qiime2 command: first example



Sample01_S10_L001_R1_001.fastq
Sample01_S10_L001_R2_001.fastq
Sample02_S11_L001_R1_001.fastq
Sample02_S11_L001_R2_001.fastq



```
qiime tools import --type 'SampleData[PairedEndSequencesWithQuality]' \
--input-path ./reads/ \
--source-format CasavaOneEightSingleLanePerSampleDirFmt \
--output-path ./imported_reads.qza
```



imported_reads.qza

Qiime2 command: first example



imported_reads.qza

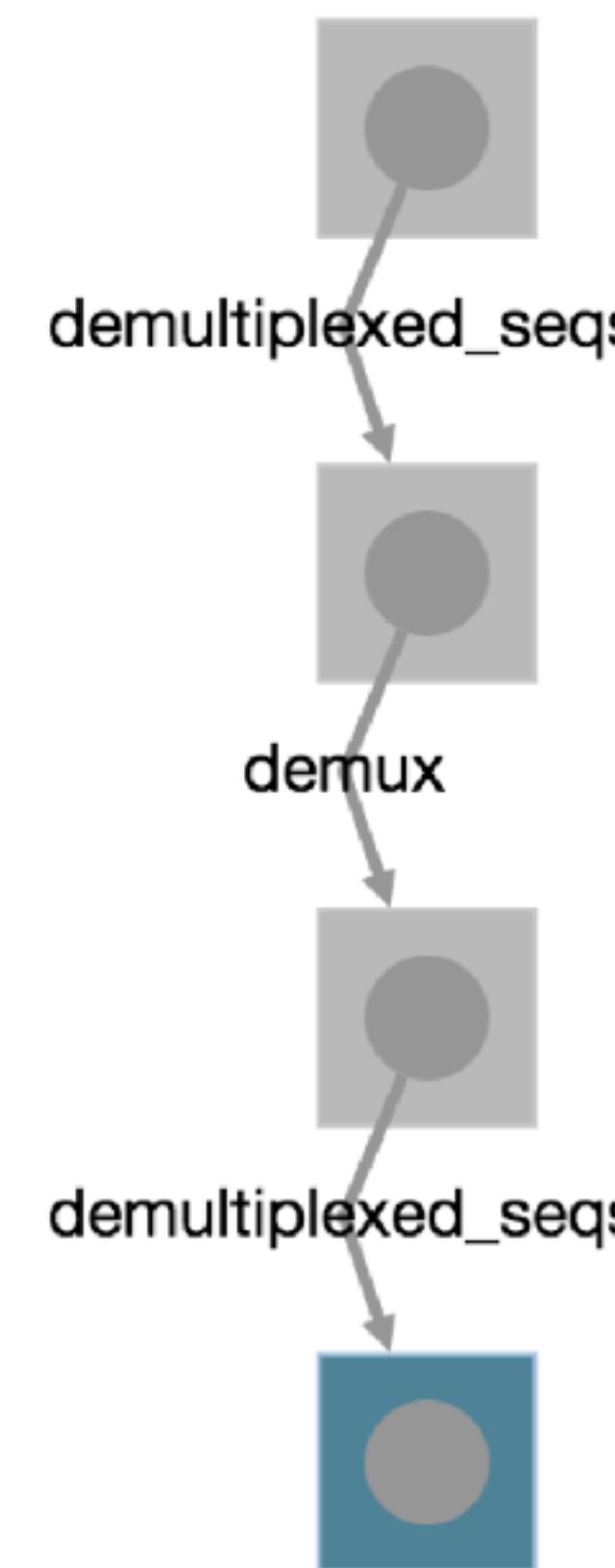
```
$ unzip -t demux-paired-end.qza
```

```
Archive: demux-paired-end.qza
testing: e2150a41-6d7c-4e13-99ee-36f57ab1f2fb/VERSION      OK
testing: e2150a41-6d7c-4e13-99ee-36f57ab1f2fb/metadata.yaml  OK
testing: e2150a41-6d7c-4e13-99ee-36f57ab1f2fb/provenance/VERSION  OK
testing: e2150a41-6d7c-4e13-99ee-36f57ab1f2fb/provenance/metadata.yaml  OK
testing: e2150a41-6d7c-4e13-99ee-36f57ab1f2fb/provenance/citations.bib  OK
testing: e2150a41-6d7c-4e13-99ee-36f57ab1f2fb/provenance/action/action.yaml  OK
testing: e2150a41-6d7c-4e13-99ee-36f57ab1f2fb/data/64_S64_L001_R2_001.fastq.gz  OK
testing: e2150a41-6d7c-4e13-99ee-36f57ab1f2fb/data/34_S34_L001_R1_001.fastq.gz  OK
testing: e2150a41-6d7c-4e13-99ee-36f57ab1f2fb/data/60_S60_L001_R2_001.fastq.gz  OK
testing: e2150a41-6d7c-4e13-99ee-36f57ab1f2fb/data/19_S19_L001_R1_001.fastq.gz  OK
```

“Provenance” metadata



Provenance Graph



Action Details

▼ execution:
 uuid: "4ccc8855-337b-4db6-9715-c3dd59f86251"
▼ runtime:
 start: 2018-03-19T08:59:26.488Z
 end: 2018-03-19T15:22:47.725Z
 duration: "6 hours, 23 minutes, 21 seconds, and 236906 microseconds"
▼ action:
 type: "method"
 plugin: "environment:plugins:deblur"
 action: "denoise_16S"
▼ inputs:
 ▼ 0:
 demultiplexed_seqs: "2c0cf238-bbe9-4f0c-81d3-201144a96e03"
▼ parameters:
 ▼ 0:
 trim_length: 450

Mothur

- a single package mainly developed by Pat Schloss
- has by default an “R-like” interface
- No charts produced, but some basic numerical ecology
- Can be slow with large datasets

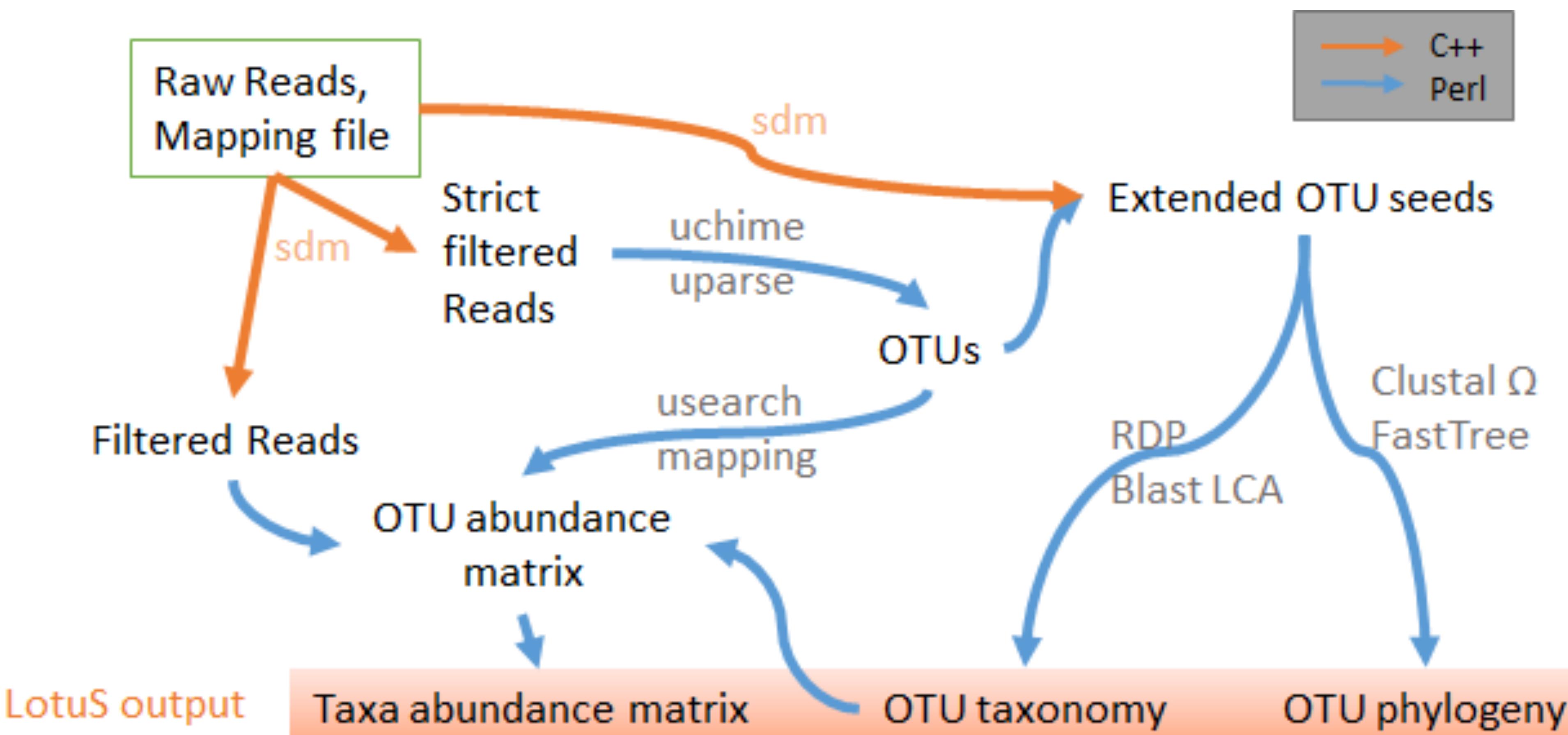


Lotus

See workshop!

- developed by Falk Hildebrand and collaborators
- designed with speed and ease of use in mind
- extensive quality filtering and demultiplexing options provided by **sdm**
- Robust LCA algorithm to assign taxonomy
- Analysis of R1 and seed extension later (quality strategy)
- Simple output folder with OTUs, OTU table and taxonomy table

Lotus

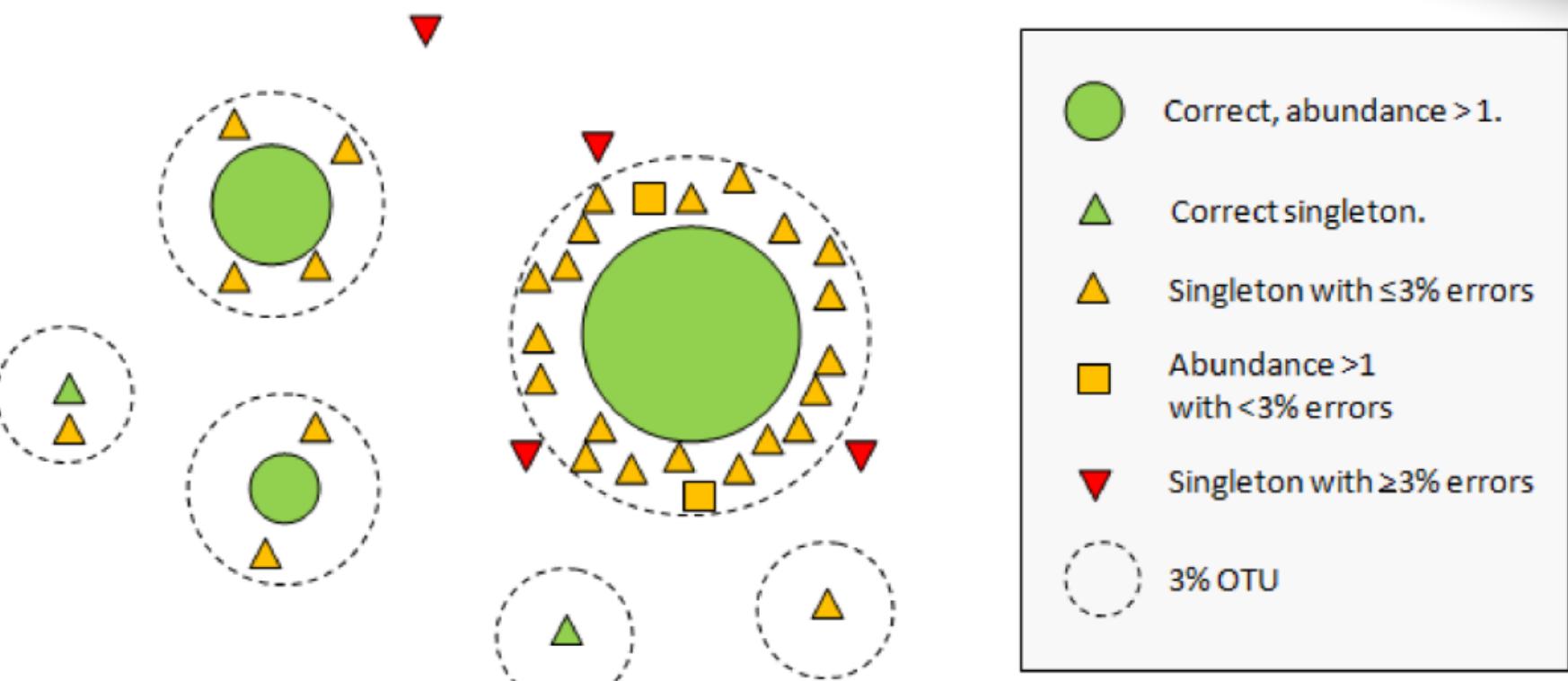


Lotus

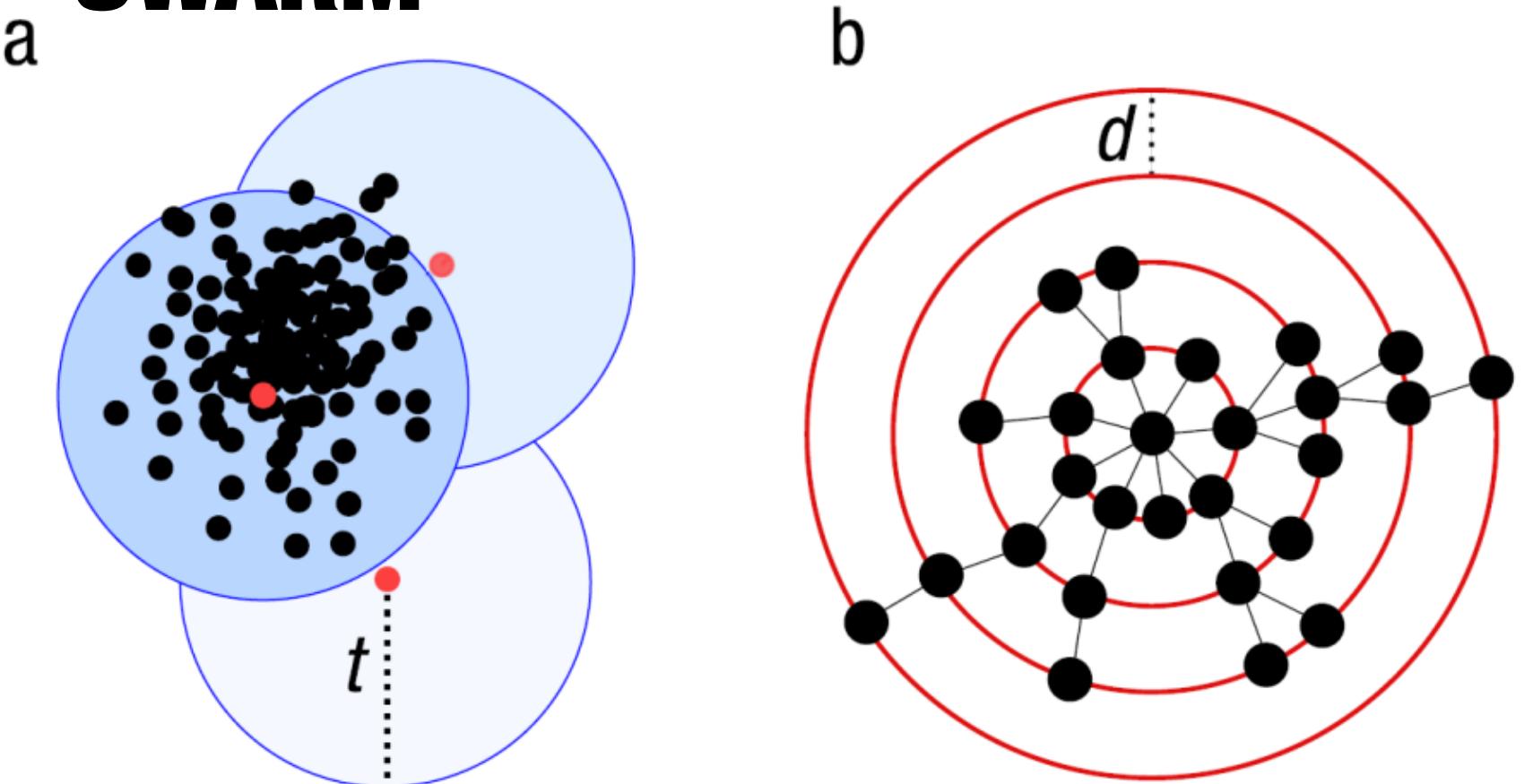
- Clustering can be done with **UPARSE**, **CD-HIT** or **SWARM**
- Taxonomy: multiple DBs and algorithms (RDP, BLAST, LAMBDA)
- Ships a fast quality filtering and demultiplexing program (**sdm**): clustering input is a set of clean seqs

See workshop!

USEARCH



SWARM



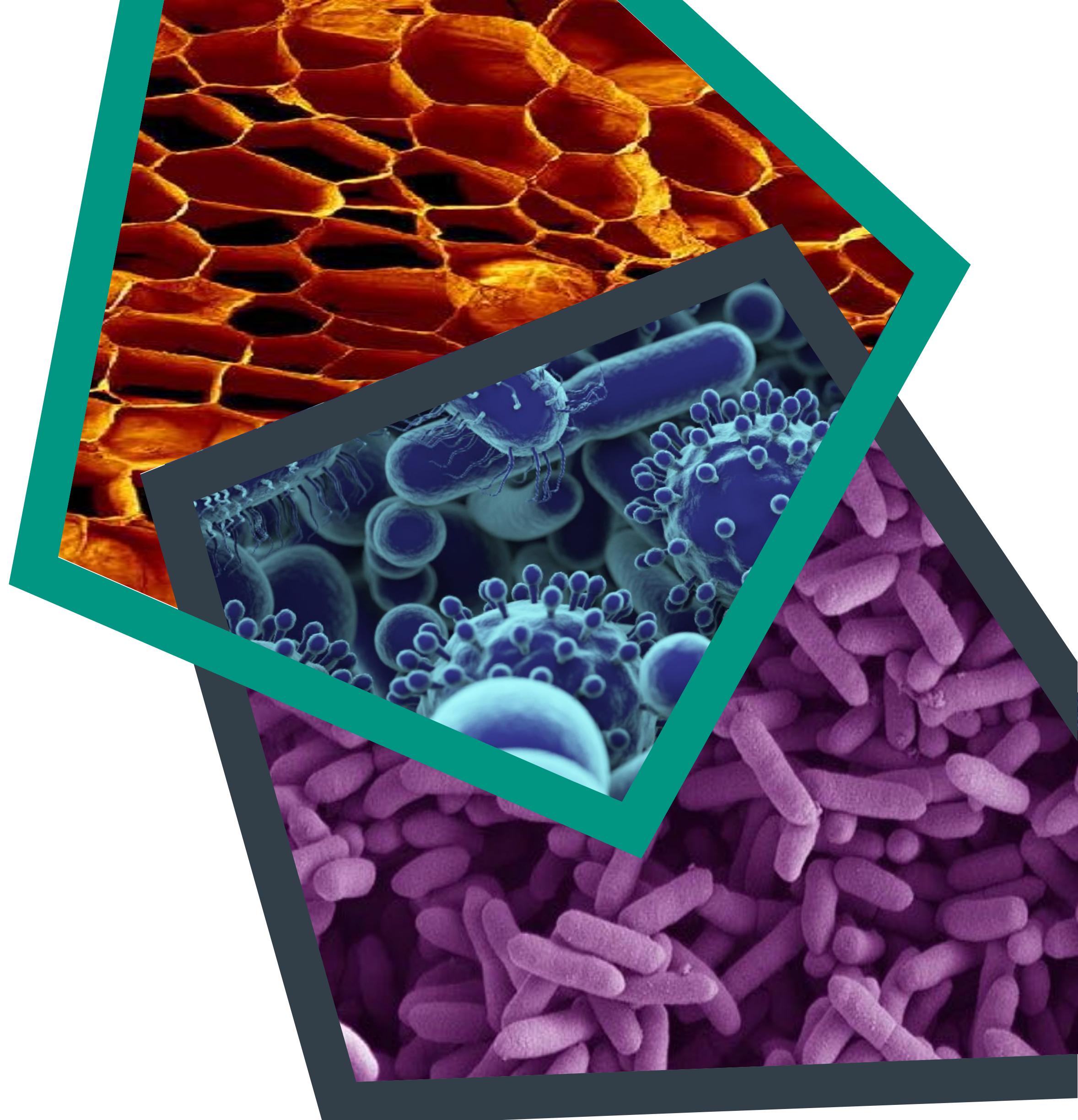


Science ▾ Health ▾
Food ▾ Innovation

Hands on Workshop



<https://github.com/telatin/lotus-tutorial>



Metadata

A “Qiime” like Mapping File

Input Formats

“sdm” can be used to
demultiplex and quality filter

Sequence
Analysis

Lotus



Metadata

It's a TSV file with a fastq with the **fastqFile** field linking to the input files.

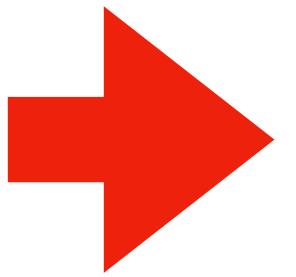
#SampleID	Facility	fastqFile
13xSPFxCD	SPF	13xSPFxCD_R1.fastq,13xSPFxCD_R2.fastq
14xSPFxCD	SPF	14xSPFxCD_R1.fastq,14xSPFxCD_R2.fastq
15xSPFxCD	SPF	15xSPFxCD_R1.fastq,15xSPFxCD_R2.fastq
16xSPFxCD	SPF	16xSPFxCD_R1.fastq,16xSPFxCD_R2.fastq
19xSPFxHFD	SPF	19xSPFxHFD_R1.fastq,19xSPFxHFD_R2.fastq
21xSPFxHFD	SPF	21xSPFxHFD_R1.fastq,21xSPFxHFD_R2.fastq
22xSPFxHFD	SPF	22xSPFxHFD_R1.fastq,22xSPFxHFD_R2.fastq
24xSPFxHFD	SPF	24xSPFxHFD_R1.fastq,24xSPFxHFD_R2.fastq

Metadata

- Lotus provides a script to generate a minimal mapping file (autoMap.pl), but it will only contain the paths and samples will have progressive names (e.g. SMPL1, SMPL2...)
- Its output can be edited (also with Excel) to add the extra columns and then copied and pasted in text format with tabs
- Can be checked `lotus.pl -check_map MAPPING.TXT`

#SampleID	Facility	fastqFile
13xSPFxCD	SPF	13xSPFxCD_R1.fastq,13xSPFxCD_R2.fastq
14xSPFxCD	SPF	14xSPFxCD_R1.fastq,14xSPFxCD_R2.fastq

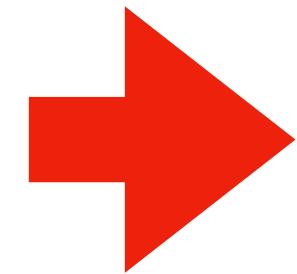
Input Formats



sdm

- Common scenario: a set of FASTQ files demultiplexed (Illumina MiSeq, Illumina HiSeq...)
- Demultiplexing: you have an Illumina run with the barcodes (indexes) in a separate file: **sdm can demultiplex**
- Old datasets: you download a .fna and .qual file of a 454 Run, that had the barcode (MID) at the beginning of the sequence:
sdm can demultiplex

Analysis



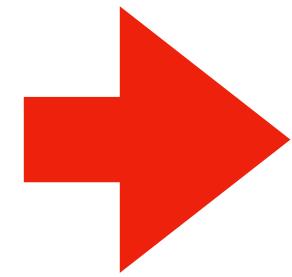
lotus

- Basic usage:

```
lotus -i input_folder -m mapping.txt \
-o output [-c conf.file -s sdm.file]
```

- The **configuration** file allows to set several options, being a “protocol” file. There is a template file in the lotus installation dir
- Since Lotus can handle demultiplexing, you can supply demux and quality filtering parameters with an sdm file,

Analysis



lotus

- Full documentation: [link to lotus.md](#)
- `-simBasedTaxo [blast, lambda...]` to change classification algorithm (default: RDP)
- `-refDB [GG, SLV...]` to change reference database
 - a custom database can be used (see tutorial)